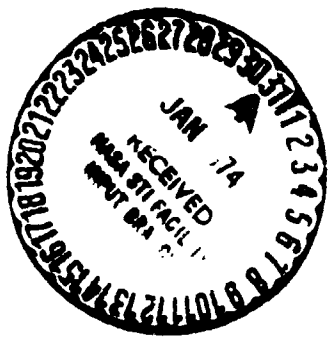
NASA TECHNICAL MEMORANDUM

NASA TM X-64801



NOTES FOR THE IMPROVEMENT OF THE SPATIAL AND SPECTRAL DATA CLASSIFICATION METHOD

By Charles C. Dalton
Aero-Astrodynamics Laboratory

January 23, 1974

(NASA-TM-X-64801) NOTES FOR THE IMPROVEMENT OF THE SPATIAL AND SPECTRAL DATA CLASSIFICATION METHOD (NASA) 33 p HC $3.75 CSCL 05B

N74-15058

Unclas
G3/13 26826

NASA

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
NASA TM X-64801		

4. TITLE AND SUBTITLE	5. REPORT DATE
Notes for the Improvement of the Spatial and Spectral Data Classification Method	January 23, 1974
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT #
Charles C. Dalton	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NO.
George C. Marshall Space Flight Center Marshall Space Flight Center, Alabama 35812	11. CONTRACT OR GRANT NO.
	13. TYPE OF REPORT & PERIOD COVERED
12. SPONSORING AGENCY NAME AND ADDRESS	Technical Memorandum
National Aeronautics and Space Administration Washington, D. C. 20546	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This work was conducted as part of the Aerospace Environment Division, Aero-Astrodynamic Laboratory Laboratory, program to develop environmental data management and classification techniques applicable to remote sensing measurements.

16. ABSTRACT

This report examines the "Spatial and Spectral Clustering" technique for the unsupervised automatic classification and mapping of earth resources satellite data, and makes theoretical analysis of the decision rules and tests in order to suggest how the method might best be applied to other flight data such as Skylab and Spacelab.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
Unsupervised Classification Earth Resources Mapping ERTS Algorithm Pattern Recognition Spatial and Spectral Clustering	Unclassified-unlimited William J. Winfrood Ludie G. Richard, Acting Director Aero-Astrodynamics Laboratory

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	32	NTIS

TABLE OF CONTENTS

Notes for the Improvement of the Spatial and Spectral Data Classification Method

I. INTRODUCTION

A. Background

In the corresponding part of a recent report by the author[1] a detailed explanation was given for the present interest in non-supervised techniques for the automatic classification of satellite multispectral ground scene data vis a vis the techniques involving supervised computation. A familiarity with Jayroe's[2] report in this area is also presupposed.

B. Present Situation

The author was asked to make a theoretical evaluation of Su's[3] and Jayroe's[2] quite different approaches to non-supervised classification of satellite multispectral ground scene data. The author chose to do that effort in three separate steps: (1) to evaluate Su's[3] model first independently of Jayroe's[2] model and to suggest any likely improvements which would retain the same general idea of the approach, (2) to do the same for Jayroe's[2] model, and (3) after seeing the effects of the changes by processing some data with the resulting revised algorithms, to propose what new model might combine the best compatable parts or compromises from the two models. The first step was covered in References 1 and 4. Reference 4 gives the complete algorithm, which was given only for the first pass of the data in Reference 1, and which is included herein as Appendix A.

II. DISCUSSION

A. Description

Jayroe's[2] unsupervised feature extraction process was developed for the analysis of flight data which has n spectral channels responding to each elemental area of the ground-scene which is resolved in rectangular coordinates x and y. His method has four stages, which one can describe briefly as follows:

(1) A boundary map of the data is produced by separating the data into homogeneous and inhomogeneous areas. Each resolution element has a root mean square spectral difference s_x or s_y with respect to the element which is adjacent to it in the x or y direction. Any element where s_x or s_y is equal to or less than the average of such values for all of the elements in the scene is classified as a homogeneous element; otherwise, the element is classified as a boundary. A digital image of a boundary map is recorded on magnetic tape for use in the next stage of processing. See Section II. B. 1 for comments.

(2) The second stage is concerned with the selection and spatial merging of unknown candidate features based upon the homogeneity of the ground scene, as displayed by the boundary map which was recorded on magnetic tape in the first stage. See Section II. B. 2 for comments.

(3) The third stage of processing is concerned with spectral merging of the selected unknown candidate features. In this stage the decision, to merge or not to merge, is based entirely upon spectral information rather than the spatial information which was used in the

second stage. The boundary and cluster map tape gives the locations of the raw data on the raw data tape belonging to each cluster. The mean feature vectors and covariance matrices are calculated for each cluster. "These calculations are used to define decision boundaries with which to physically surround the data belonging to a cluster in n-dimensional space. The most general closed surface that can be used to surround the n-dimensional data is an n-dimensional hyperellipse. The centroid of the cluster ellipse is given by the feature vector mean values $_k\bar{x}_\ell$ ..." (quoting Jayroe[2]). One makes "... a rotation, E_ℓ, followed by a diagonal transformation, W_ℓ ..." "Thus, the equation of an n-dimensional ellipse in reduced form is obtained for each cluster, and, in general, each cluster will have a different coordinate system. The next step is to give a decision rule for determining how many clusters actually represent the same feature... The decision rule is that two clusters represent the same feature if the centroids of both clusters are contained in both clusters' ellipses." See Section II. B. 3 for comments and analysis.

(4) Jayroe[2] explains that the final stage of processing is concerned with classifying the data in the digital image of the ground scene and with showing the location and distribution of the features. The inputs to this stage of processing are the raw data tape, the statistics for each class, and the boundary tape. The decision rule which Jayroe[2] chose for classifying a resolution element into a given class, and the basis which he gave for it, are discussed with some analysis in Section II. B. 4 herein.

B. Analysis and Evaluation

1. Stage One: Boundary Mapping

Jayroe[2] considers the equation of an ellipse in the (s_x, s_y) plane, involving quadratic and product terms. By using all of the resolution elements in the ground scene he finds the sample mean values of s_x^2, s_y^2, and $s_x s_y$. Those values are used to determine what transformation will align the coordinates with the principal axes and give the values of the semi-major and semi-minor axes of the particular ellipse which the sample mean values infer. That particular ellipse is then found in the (s_x, s_y) coordinates after the inverse transformation, from which the values of a, b, and c are determined when the sample mean ellipse is

$$as_x^2 + bs_y^2 + cs_x s_y = 1. \qquad (1)$$

One could then say, as Jayroe[2] does, that the decision is to classify a resolution element as being homogeneous unless the left side of equation (1) exceeds unity. Or, maybe one should say that the left side of equation (1) is a random variable such that the sample estimator of its mean is unity, and that the decision is to classify a resolution element as being homogeneous if

$$as_x^2 + bs_y^2 + cs_x s_y \leq B \qquad (2)$$

where B is an adjustable parameter. One could give B a higher value than unity as a trade-off against excessive computer time, up to some maximum value of B beyond which experience would show that boundary formation would be dampened enough to reduce effectiveness materially.

2. Stage Two: Cluster Formation

Whereas, the first stage identified each resolution element of the ground scene as being either a boundary element or a homogeneous element, it became of interest to see what is the smallest number of elements which a cluster or class could have. It seems that this depends on the second stage. Clusters of homogeneous elements are formed in the second stage, and the resulting clusters are merged into classes in the third stage. The fourth stage then classifies each element as belonging to or not belonging to the established classes. Thus, every cluster and every class has at least p^2 members; i.e., no element can be classified unless it is sufficiently nearly like those which cluster in a homogeneous area which extends beyond a square array containing p^2 elements. Jayroe[2] suggests 100 elements for the pxp array. In contrast to this the models by Su[3] and by Dalton[1,4] can classify any isolated element which is sufficiently nearly like any five other elements (which do not even have to be together). The resolution elements in the ERTS data are each about 79.2 meters (1/20 mile) x 57.2 meters; i.e., in practical terms, a square field of less than about 5/8 square kilometer (one quarter section) would not accept the 10 x 10 array, and larger fields are usually not entirely homogeneous.

3. Stage Three: Spectral Merging

a. Decision Rule

In the computations for n spectral channels, Jayroe[2] made transformations (a rotation F_ℓ followed by a diagonal transformation W_ℓ) to reduce to canonical form the covariance matrix for each of the

clusters of homogeneous ground scene elements. This may be computationally efficient; also, it facilitates theoretical derivations for decision rules because the transformed dimensions become statistically indepedent. Jayroe's[2] analysis through his equation (26) is verified. Instead of Jayroe's[2] equation (28) for the inverse similarity S^{-1} for the cluster ℓ and the vector v_k (which is the mean which k's transformed coordinate system gave for cluster k) as viewed within 1's coordinate system, one gets

$$S^{-1} = \left(\prod_{p=1}^{n} {}_{pp}c_{\ell}^{1/2} \right)^{2/n} \left[2n + \sum_{p=1}^{n} \frac{\left({}_{p}v_{k} - {}_{p}\bar{x}_{\ell} \right)}{{}_{pp}c_{\ell}} \right] \qquad (3)$$

when the c's are variances and where the term 2n, instead of just n, corrects for a term ${}_{pp}c_{k}$ which Jayroe inadvertently omitted from the bracketed factor in his equation (27).

For the expected value of S^{-1} in this equation (3) Jayroe[2] just replaced the summation by n. That result would seem to be due to (1) an oversight in which the v_k may have been considered to represent an individual resolution element as a prospective member of the cluster ℓ, whereas it is instead the mean of an entire cluster k, followed by (2) an assumption that the summation would have approximately a chi-square distribution with n degrees of freedom, and further (3) an assumption that the coefficient in equation (3) does not vary appreciably.

Attention will be given to the points mentioned about S^{-1} in equation (3). First, though, it may be recognized as an expedient departure from rigor in that the n-dimensional space for n spectral channels of data has for each cluster a separate transformation. Yet, the practical objective which Jayroe[2] pursues is to reduce computational requirements by sufficiently nearly achieving statistical independence between the n terms of the summation in equation (3). This approach seems to be so nearly a characteristic to be proven by the results that it is retained as a constraint on the present analysis.

Dalton's[1] analysis considered that, when normal basic variables ${}_p x_k$ and ${}_p x_\ell$ have the same population mean, $F_{p,\,k\ell}$ or $t^2_{p,\,k\ell}$ has an F distribution with one and $M_k + M_\ell - 2$ degrees of freedom when (in the same coordinate system) M_k and M_ℓ random samples of classes k and ℓ have means ${}_p\bar{x}_k$ and ${}_p\bar{x}_\ell$ and variances ${}_{pp}c_k$ and ${}_{pp}c_\ell$:

$$F_{p,\,k\ell} = t^2_{p,\,k\ell} \tag{4}$$

$$= \frac{M_k M_\ell (M_k + M_\ell - 2)\,({}_p\bar{x}_k - {}_p\bar{x}_\ell)^2}{(M_k + M_\ell)\,(M_k\ {}_{pp}c_k + M_\ell\ {}_{pp}c_\ell)} \tag{5}$$

because $t_{p,\,k\ell}$ has Student's t distribution with $M_k + M_\ell - 2$ degrees of freedom. Notice that $F_{p,\,k\ell}$ in equation (5) can be written as

$$F_{p,\,k\ell} = \left(\frac{M_k + M_\ell - 2}{4}\right)\left[\frac{\left({}_p\bar{x}_k - {}_p\bar{x}_\ell\right)^2}{\left(M_k\ {}_{pp}c_k + M_\ell\ {}_{pp}c_\ell\right)\left(\frac{1}{M_k} + \frac{1}{M_\ell}\right)/4}\right]. \tag{6}$$

Then, when the two clusters are of equal size and give equal estimators of variance ${}_{pp}c$, the bracketed term in equation (6) identifies with the term being summed in equation (3) and is otherwise also an appropriate average for ${}_{pp}c$ for the two clusters. In a similar manner the term 2n, already mentioned in equation (3), came by replacing $({}_{pp}c_k + {}_{pp}c_\ell)/{}_{pp}c_\ell$ by 2. Although equation (3) is already appropriate for use in the next stage to classify an individual prospective member of a class, symmetry seems to require for the present stage (when two clusters are to be merged) that the bracketed term in equation (6) should replace the term being summed in equation (3); i.e.,

$$S_{k\ell}^{-1} = \left(\prod_{p}^{n} {}_{pp}c_\ell^{\,1/2} \right)^{2/n} \left[2n + \left(\frac{4}{M_k + M_\ell - 2} \right) \sum_{p=1}^{n} F_{p,\,k\ell} \right] . \qquad (7)$$

The two factors in equation (7) are not statistically independent. However, they can be treated as statistically independent for the purpose of identifying parameter combination regions over which the relative variation of one of the factors is small relative to that of the other factor. The expected value $\mu_{F_{p,\,k\ell}}$ and variance $\sigma^2_{F_{p,\,k\ell}}$ of each of the n terms $F_{p,\,k\ell}$ in equation (7) are, by Reference 5,

$$\mu_{F_{p,\,k\ell}} = \frac{M_k + M_\ell - 2}{M_k + M_\ell - 4} , \quad M_k + M_\ell > 4 \qquad (8)$$

$$\sigma^2_{F_{p,\,k\ell}} = 2\mu_{F_{p,\,k\ell}}^{\,2} \left(\frac{M_k + M_\ell - 3}{M_k + M_\ell - 6} \right) , \; M_k + M_\ell > 6 . \qquad (9)$$

Computations involving pairs of clusters can be made under the transformation peculiar to either cluster in the pair, but not both transformations together. A practical expedient would be to use only the transformation determined for the larger one of the two clusters. In that case, in equations such as (5) and (6) one would replace the variance of the smaller one by the variance of the larger one of the two clusters; e.g., see equation (20). Then, the statistical independence provided by the transformation gives the mean and variance of the summation of the n terms as n times the respective values for the individual terms. Therefore, the expected value μ_{T_2} and variance $\sigma^2_{T_2}$ of the bracketed factor T_2 in equation (7) are

$$\mu_{T_2} = 2n \left(\frac{M_k + M_\ell - 2}{M_k + M_\ell - 4} \right) \tag{10}$$

$$\sigma^2_{T_2} = \frac{8n(M_k + M_\ell - 2)(M_k + M_\ell - 3)}{(M_k + M_\ell - 4)^2 (M_k + M_\ell - 6)} \quad . \tag{11}$$

Let $T_1^{2/n}$ represent the coefficient factor in equation (7) in which each of the n independent standard deviations $_{pp}c^{1/2}$ has an expected value $b_1 \sigma$ and a variance which, by Reference 5, can be taken as approximately $\sigma^2/2M_\ell$. Then n products of such statistically independent factors has a mean μ_{T_1} which is

$$\mu_{T_1} = (b_1 \sigma)^n \quad . \tag{12}$$

Within the accuracy of a first order theory for the propagation of error, the variance of T_1 can be approximated by

$$\sigma^2_{T_1} \simeq \frac{T_1^2 \quad n \quad (\sigma^2/2M_\ell)}{(b_1 \sigma)^2}$$

$$\simeq \frac{nT_1^2}{2b_1^2 M_\ell} \quad . \tag{13}$$

But if the expected value of T_1 is $(b_1 \sigma)^n$ in equation (12), what can one best say might be the expected value of $T_1^{2/n}$? Of course it would depend on the distribution function. With ERTS data, the value of n is 4; so one wants the expected value of a square root. Also, one knows that the result $T_1^{2/n}$ is the geometric mean of the variances of the basic variables. Therefore, it may be sufficiently accurate to approximate it by the 2/n power of $(b_1 \sigma)^n$, which is $(b_1 \sigma)^2$. The variance of $T_1^{2/n}$, by a further application of first order theory of error propagation, is

$$\sigma^2_{T_1^{2/n}} \simeq \left(\frac{2}{n} T_1^{2/n}/T_1\right)^2 \sigma^2_{T_1}$$

$$\simeq \frac{2b_1^2 \sigma^4}{nM_\ell} \tag{14}$$

where values of b_1 as a function of M_ℓ are tabulated in Reference 5 (wherein b_1 is called b(n) and M_ℓ is called n).

$$\mu_{T_1^{2/n}} \simeq b_1^2 \sigma^2 \tag{15}$$

Thus, to the extent that the two factors for S^{-1} in equation (7) can be considered statistically independent, the expected value of S^{-1} is

$$\mu_{S^{-1}} = \mu_{T_1^{2/n}} \mu_{T_2}$$

$$\simeq 2(b_1 \sigma)^2 \, n \left(\frac{M_k + M_\ell - 2}{M_k + M_\ell - 4}\right). \tag{16}$$

Then, by a further application of the first order approximation of error propagation, the variance of S^{-1} is, relatively,

$$\sigma^2_{S^{-1}} = \mu^2_{T_2} \sigma^2_{T_1^{2/n}} + \mu^2_{T_1^{2/n}} \sigma^2_{T_2}$$

$$\simeq \frac{8b_1^2 n\sigma^4 (M_k + M_\ell - 2)}{(M_k + M_\ell - 4)^2} \left[\left(\frac{M_k + M_\ell - 2}{M_\ell}\right) + b_1^2\left(\frac{M_k + M_\ell - 3}{M_k + M_\ell - 6}\right)\right]. \tag{17}$$

Thus, the expected value and variance of S^{-1} are both proportional to n, but the ratio of the relative contributions to the variance of S^{-1} due

to the two factors in equation (7) is independent of n. When the two clusters k and ℓ are the same size, then the second factor in equation (7) makes a contribution to the total variance which decreases from 45 percent for two clusters of size sever each, but remains very nearly a constant 1/3 of the total variance for any cluster size of 10 or more. So, the assumption of statistical independence between the two factors in equation (7) would seem to be problematical for getting any accurate estimate of the expected value of S^{-1}, etc., which Jayroe[2] pursued in his equation (29). However, there seems to be no easy alternative to choosing some reasonable approximation to a decision rule which would merge two clusters if the summation in equation (7) would not exceed its expected value plus the product of some parameter C (which may be a constant or a function of n, see the last paragraph in this section) and the theoretical value of the standard deviation of the summation; i.e.,

$$\sum_{p=1}^{n} F_{p,\,k\ell} \leq n\,\mu_{F_{p,\,k\ell}} + C\sqrt{n}\;\sigma_{F_{p,\,k\ell}} \tag{18}$$

where $F_{p,\,k\ell}$ is given by equation (5), $\mu_{F_{p,\,k\ell}}$ is given by equation (8), and $\sigma_{F_{p,\,k\ell}}$ is the square root of the variance in equation (9). Then, by substituting from the cited equations into equation (18) and rearranging the material, one gets

$$\left(\frac{M_k + M_\ell - 4}{M_k + M_\ell}\right) \sum_{p=1}^{n} \frac{(\bar{x}_{pk} - \bar{x}_{p\ell})^2}{\frac{C_{pk}}{M_\ell} + \frac{C_{p\ell}}{M_k}} \leq n + C\left(\frac{M_k + M_\ell - 3}{M_k + M_\ell - 6}\right)^{1/2} \sqrt{2n} \tag{19}$$

where, yet rigorously, the terms on the right side are the expected value and C standard deviations of the left side. One can now propose that, in equation (19), the factors involving the sums of the sample sizes might should be eliminated as a practical expecient. The elimination of the factor on the left side would account for most of the error, would be of no practical consequence for large clusters; it would cause an error in the mean of only 4 percent when the sum of the two clusters is 100, and this would not exceed 1/7 of the standard deviation when the number of channels n does not exceed 24. In practice, the given example is not intended to suggest any such size as a lower limit for cluster size; most clusters are larger than any permitted minimum size, and typically the combined size of two clusters is considerably larger than twice any minimum size. Therefore, as a more practical expedient than the more rigorous equation (19), an appropriate decision rule would seem to be that two clusters or classes k and ℓ of sizes M_k and M_ℓ should be combined into the same class when

$$\sum_{p=1}^{n} \frac{(p\bar{x}_k - p\bar{x}_\ell)^2}{\frac{pp C_k}{M_\ell} + \frac{pp C_\ell}{M_k}} \leq n + C\sqrt{2n}.$$

Let the designations of the clusters k and ℓ be such that $M_\ell \geq M_k$. Then, by the expedient which was discussed in the paragraph following equation (9), by replacing $_{pp}C_k$ with $_{pp}C_\ell$ and using the transformation

determined for cluster ℓ in the summation on the left side, the decision rule becomes

$$\left(\frac{M_k M_\ell}{M_k + M_\ell}\right) \sum_{p=1}^{n} \frac{(p\bar{x}_k - p\bar{x}_\ell)^2}{pp C_\ell} \leq n + C\sqrt{2n}. \qquad (20)$$

Jayroe's[2] results, both in his equation (29) for the expected value of S^{-1} and in his equation (30) decision rule for merging two clusters, would seem to require that the denominator in each term in the summation in equation (3) (Jayroe's[2] equation (28)) would be the variances with respect to the means instead of the variances of the variables before they are averaged. Therefore, it would seem that the criterion which Jayroe[2] has in his equation (30) exceeds the expected value of the indicated summation by a factor which would be approximately half of the size of a cluster. One would expect that the model in that form might show a tendency to combine clusters excessively.

Jayroe[2] notes that the decision rule, his equation (30), is a hyperellipse in the principal axis coordinates; that seems yet to be true with equation (20). He says that the threshold in the decision rule (the right side of the equation) is independent of the cluster and depends only on the dimension n of the feature space; that is true also in equation (20). However, Jayroe[2] added: "Thus, if an elliptical boundary decision rule is used in the principal axis coordinate system, the theorem can be extended to say that the diagonal transformation is not needed and only the eigenvector transformation is

needed since the threshold can always be written as some constant times [the geometric mean of the n variances of cluster ℓ]," which is the coefficient factor $T_1^{2/n}$ in equations (3) and (7) with its expected value and variance approximated in equations (15) and (14), respectively. Unfortunately, though, both the expected value and the variance of the cited function depend on the unknown variance σ^2 of the population of which the given cluster is only a random sample of size M_ℓ. It is agreed that the diagonal transformation is not needed for computations; it does, however, show the origin and context of the equation (20) decision rule. The sample means and variances which are used in the summation on the left side of equation (20) are given by Jayroe's[2] equations (17) and (18), without the diagonal transformations; they do, though, presuppose that the computations are done in principal axis coordinates in order that the n terms in the summation in equation (20) are statistically independent.

The principal axis of clusters which are random samples from the same population will have some distribution with respect to the principal axis of the population. Thus, the principal axis for cluster k will generally be different from those for cluster ℓ, and different from those which follow from combining the two clusters. It is expected that it will be sufficiently accurate to use the computational expedient which ignores the distinction cited because the hypothesis being tested by the equation (20) decision rule is that the two clusters are from the same population.

Jayroe's[2] model involves two determinations of his decision rule, his equation (30), by reversing the roles of the two clusters because that equation is not symmetric with respect to the two

clusters k and ℓ. The proposed revised decision rule, equation (20), will require only half as much computation because it computes a transformation for only the largest one of the two clusters.

In Jayroe's[2] model, using his equation (30) as a decision rule for combining clusters or classes, it is not likely that a sufficient number of classes would tend to remain for some purposes. With equation (20), however, the number of clusters or classes which will remain distinct will depend on the value chosen for the adjustable parameter C. Three considerations are evident: (1) all clusters which represent the same population class should be merged, (2) unlike classes should not be merged except, (3) when there are more statistically distinct classes than some upper limit which must be imposed as a computational or other constraint, then further merging is necessary. The statistical significance of values of C, except for the smaller clusters, is illustrated approximately by: (1) a value of C of $-(1 + n/100)$ would combine all pairs of clusters which show less than a 10 percent confidence level of being from different populations, (2) a value of C of $-(2/3)/\sqrt{2n}$ would combine no clusters which differ by more than a 50 percent confidence level, and (3) a value of C of 4/3 would combine all clusters except those which show at least a 90 percent confidence of distinct populations.

b. Order of Merging

In Jayroe's[2] model the distance between cluster centers was not considered in choosing which pair of clusters should be tested for merging. It would seem that the order of merging would

effect the quality of the results. Some of the computation time which is saved by using equation (20) can well be expended toward this improvement. When, in the course of the analysis, there remain K clusters or classes then there corresponds a square symmetric matrix of center separation values which upon being ranked have some smallest value, possibly repeated. The corresponding pair of clusters k and ℓ should be tested by equation (20) to see if they should be merged. If they are merged, then in the matrix columns k and ℓ and rows k and ℓ are deleted and are replaced by one new row and column. The smallest value is again sought, etc. But if clusters k and ℓ are not merged, then their element in the matrix is replaced by a number larger than the largest element before proceeding, and another matrix of uncombinable pairs is begun whose elements are the values of the left side of equation (20), etc. Another decision rule will be needed so that when the distance between centers exceeds a certain value the equation (20) test will be skipped

$$\sum_{p=1}^{n} (\bar{x}_{pk} - \bar{x}_{p\ell})^2 \leq E_n .$$

Then, if the number of remaining clusters or classes exceeds the maximum allowable number, any further reduction is made by using the matrix of computed values of the left side of equation (20), so far as it had been used; the smallest element would identify the pair to be merged even though they qualified as distinct classes.

4. Stage Four: Classification

The decision rule which Jayroe[2] uses, his equation (31), for deciding when an individual element can be added to a particular class in the final classification is that the summation in equation (3)

must not exceed 2n. The explanation that the factor 2 is used because the exponent in a normal distribution is divided by that factor is not convincing. As he says, though, it does seem appropriate to use a less restrictive criterion than that which would be right for deciding about merging two clusters. Actually, considering that the mean and variance are approximately n and 2n for large clusters, the given criterion amounts to adding $(n/2)^{1/2}$ standard deviations to the mean, which, with the 12-channel data reported[2], would classify an individual resolution element into a class unless its difference is significant at a 98 percent confidence level. With 4-channel data, as in ERTS, the confidence would be 90 percent instead of 98 percent with the same decision rule, Jayroe's[2] equation (31).

It seems prudent to derive more rigorously a decision rule which does not presuppose large clusters for classifying individual resolution elements into established clusters. For this purpose the presupposition of normal variables in principal axis coordinate systems will be continued, and the same notation as in equation (3) except that $_pv$ is the coordinate of an individual resolution element instead of the sample mean of a cluster. Dalton[1] showed that

$$\left(\frac{M_\ell - 1}{M_\ell + 1}\right) \sum_{p=1}^{n} \frac{(_pv - _p\bar{x}_\ell)^2}{_pp^C_\ell} = \sum_{p=1}^{n} F_{p,\,\ell}\,(1,\, M_\ell - 1) \qquad (21)$$

where each of the n terms on the right has an F distribution with one and M_ℓ-1 degrees of freedom, and they are statistically independent due to the principal axes coordinates. Then, for $M_\ell \geq 6$ this gives, in the

form of a decision rule as a function less than or equal to its expected value plus the product of some parameter D and the standard deviation,

$$\left(\frac{M_\ell - 3}{M_\ell + 1}\right) \sum_{p=1}^{n} \frac{(p^v - p^{x_\ell})^2}{pp^{C_\ell}} \leq n + D\sqrt{2n}\left(\frac{M_\ell - 2}{M_\ell - 5}\right)^{1/2} \quad .(22)$$

5. Further Passes

Jayroe[2] explains that his program has the capability, when the size of the pxp array for cluster selection has caused incomplete classification of the ground scene, to reduce the size of the array in order to search for further clusters and to make a further classification of the data. It would appear that this is better than using a smaller array in the first place. This is because, as Jayroe[2] says, "The fixed-shape array, if chosen large enough, will not permit the mixing of features because the open gaps in the boundaries will be so small compared to the array size that the array will not be able to pass through the boundary." Jayroe's[2] statement about the 10 x 10 array, that the minimum sample size which it provides (100) is very adequate for statistical calculations, inadvertently may give the impression that a smaller array would give a sample size which would statistically not be adequate for the determination of (1) whether or not two such clusters should be merged or (2) whether or not the class which it might represent should also contain a particular individual resolution element which is to be classified. However, equations (21) and (22) are based on the F distributions instead of the chi-square distribution, the statistical requirement for which is that the clusters must not have less than 6 members.

In either case the basic variable is presupposed to be distributed approximately normally. The statistical reason that larger clusters are needed when one does not use the F distributions is that, in order to use the chi-square distribution for the summation on the left sides of equations (21) and (22), the indicated means and variances must be presupposed to be identical with those of the (unknown) population of which the cluster is a random sample of size M_ℓ.

C. Conclusions and Recommendations

Jayroe's[2] decision criterion to classify a resolution element as being homogeneous is (his equation (16))

$$as_x^2 + bs_y^2 + cs_x s_y \leq 1 \qquad (1)$$

and that otherwise the element is a boundary. It seems likely that the criterion could be improved by writing it as

$$as_x^2 + bs_y^2 + cs_x s_y \leq B \qquad (2)$$

and experimentally checking whether some other value of B in the vicinity of 1 might give a model which would have a better balance between effectiveness and computation requirement.

The decision rule which Jayroe[2] uses to see if clusters k and ℓ should be merged, when the clusters have individual transformations giving statistical independence of their n-channel spectral data, is (his equation (30))

$$\sum_{p=1}^{n} \frac{(p^{v}_{k} - p^{\bar{x}}_{\ell})^2}{pp^{C}_{\ell}} \leq n \qquad (23)$$

provided that the equation is also satisfied when the roles of k and ℓ are reversed. Instead, a better decision rule would seem to be, where k and ℓ are such that $M_{\ell} \geq M_{k}$:

$$\left(\frac{M_k M_{\ell}}{M_k + M_{\ell}}\right) \sum_{p=1}^{n} \frac{(p^{\bar{x}}_{k} - p^{\bar{x}}_{\ell})^2}{pp^{C}_{\ell}} \leq n + C\sqrt{2n} \qquad (20)$$

where M_k and M_{ℓ} are the sample sizes of class (or cluster) k and ℓ, and where C is the number of standard deviations from the expected value of the left side. Although the best value for C might be in the vicinity of zero, some experiments with data might show a better value. The clusters with the closest centers should be tested for merging before testing more distant clusters.

The decision rule which Jayroe[2] uses to see if an individual resolution element should be added to a class ℓ is (Jayroe's[2] equation (31))

$$\sum_{p=1}^{n} \frac{(p^{v} - p^{\bar{x}}_{\ell})^2}{pp^{C}_{\ell}} \leq 2n. \qquad (24)$$

Instead, a better decision rule would seem to be

$$\left(\frac{M_{\ell} - 3}{M_{\ell} + 1}\right) \sum_{p=1}^{n} \frac{(p^{v} - p^{\bar{x}}_{\ell})^2}{pp^{C}_{\ell}} \leq n + D\sqrt{2n}\left(\frac{M_{\ell} - 2}{M_{\ell} - 5}\right) \qquad (22)$$

where D is the number of standard deviations from the expected value of the left side. Because his sample sizes were sufficiently large for the purpose, Jayroe's[2] choice in equation (24) corresponds to adding $(n/2)^{1/2}$ standard deviations to expected value n, and he was using 12 channels of data. This would imply a value for D of $\sqrt{6}$, but some further experiments with equation (5) might show a better value for D.

Some experimental effort is needed to establish the best combination of values of B, C, D, and p, where p is the size of the pxp array which determines the minimum size of a cluster and where B, C, and D are the model parameters in equations (2), (20), and (22).

APPENDIX A: Reference 4

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
GEORGE C. MARSHALL SPACE FLIGHT CENTER
MARSHALL SPACE FLIGHT CENTER, ALABAMA 35812

REPLY TO
ATTN OF: S&E-AERO-YF-3-73

September 19, 1973

TO: S&E-COMP-RRV/Mr. Jack A. Jones

FROM: S&E-AERO-YF/Mr. Charles C. Dalton

SUBJECT: Request for Program of Algorithm from NASA TMX-64762 and MSFC Memo S&E-AERO-YF-2-73 on Account No. 177-32-71 (Task Agreement J99)

The subject report and memo which were recently given by me offer a method for non-supervised classification and mapping of remote sensing multispectral data. A program, which please have prepared, will enable us to study the computational performance and efficiency of that method vis a vis our other methods. The subject algorithm, in somewhat further desired detail, is as follows:

ALGORITHM FOR UNSUPERVISED CLASSIFICATION USING F DISTRIBUTIONS

For each class or prospective class one needs values for the following parameters:

m = number of members in the class

$\bar{x}_k = \frac{1}{m} \sum_{a=1}^{m} x_{ka}$, class mean in each channel $k = 1, 2, \ldots, K$

$s_k^2 = \frac{1}{m} \sum_{a=1}^{m} (x_{ka} - \bar{x}_k)^2$, class variance in each channel

$Q_{k\ell} = \frac{1}{m} \sum_{a=1}^{m} [(x_{ka} - \bar{x}_k)(x_{\ell a} - \bar{x}_\ell)]^2$ each pair of channels k and ℓ

$\lambda_{k\ell} = \left(\frac{m-1}{m+1}\right)^2 Q_{k\ell}/s_k^2 s_\ell^2$ " " " " " " "

$\mu_F = \left(\frac{m-1}{m-3}\right)$

$$\sigma_F^2 = 2\mu_F^2 \left(\frac{m-2}{m-5}\right)$$

$\mu_{\Sigma F} = K\mu_F$ where K is the number of channels

$$\sigma^2_{\Sigma F} = K\sigma_F^2 + 2 \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \lambda_{k\ell}.$$

Also, for each pair of established classes i and j containing m_i and m_j members one needs values for the following parameters:

$$\mu_{F_{ij}} = (m_i + m_j - 2)/(m_i + m_j - 4)$$

$$\sigma^2_{F_{ij}} = 2\mu^2_{F_{ij}} (m_i + m_j - 3)/(m_i + m_j - 6)$$

$$\Sigma F_{ij} = \frac{m_i m_j (m_i + m_j - 2)}{m_i + m_j} \sum_{k=1}^{K} \frac{(\bar{x}_{ki} - \bar{x}_{kj})^2}{m_i s^2_{ki} + m_j s^2_{kj}}$$

$$\lambda_{k\ell,ij} = \left[\frac{m_i m_j (m_i + m_j - 2)}{m_i + m_j}\right]^2 \frac{(Q_{k\ell,i} + Q_{k\ell,j})}{(m_i s^2_{ki} + m_j s^2_{kj})(m_i s^2_{\ell i} + m_j s^2_{\ell j})}$$

$$\mu_{\Sigma F_{ij}} = K\mu_{F_{ij}}$$

$$\sigma^2_{\Sigma F_{ij}} = K\sigma^2_{F_{ij}} + 2 \sum_{k=1}^{K-1} \sum_{\ell=k+1}^{K} \lambda_{k\ell,ij}$$

$$A_{ij} = \left|(\Sigma F_{ij} - \mu_{\Sigma F_{ij}})/\sigma_{\Sigma F_{ij}}\right|$$

The two other formulas which are always used together, with a purpose which depends on what datum is substituted for the parameter x_k, are

$$\left.\begin{aligned} \Sigma F &= \frac{m-1}{m+1} \sum_{k=1}^{K} (x_k - \bar{x}_k)^2 / s_k^2 \\ \text{and} \\ A &= \left| (\Sigma F - \mu_{\Sigma F}) / \sigma_{\Sigma F} \right| \quad . \end{aligned}\right\} \tag{54}$$

Preliminary step. Is this a re-start? No: go to step 1. Yes: go to step 25.

Step 1. Read control parameters A_o, A_1, $M(\geq 6)$, W_{max}, A_F, P, and A_R.

Step 2. Read the first M samples.

Step 3. Calculate parameters for prospective class.

Step 4. With the $\bar{x}_k$, s_k^2, etc., from step 3, calculate a value of A in equation (54) for each of the M samples by using the values of x_k for that particular sample in equation (54) with the minus sign. Does the largest value of A satisfy $A < A_o$? Yes: go to step 7. No: go to step 5.

Step 5. Discard the first sample accumulated.

Step 6. Read a new sample, then go to step 3.

Step 7. Designate a new class having the parameters extant, including the class mean of the sample values of A, say $\bar{A}$.

Step 8. Does the program reach the end of the sample? Yes: go to step 19. No: go to step 9.

Step 9. Does the number of classes W satisfy $W \leq W_{max}$? Yes: go to step 12. No: go to step 10.

Step 10. Calculate class-pair parameters A_{ij} for all combinations of classes in pairs.

Step 11. Combine the two classes i and j which give the smallest pair-parameter A_{ij} and compute the single-class parameters for the resulting class, including $\overline{A}$, etc. Go to step 12.

Step 12. Read a new sample.

Step 13. By using the values of x_k from the new sample in equation (54) with the plus sign, claculate a value of A for each of the W established classes according to their given values of m, $\overline{x}_k$, s_k^2, $\mu_{\Sigma F}$, and $\sigma_{\Sigma F}^2$. Does the smallest one of the m values of A satisfy $A \leq A_1$? Yes: add the sample to that class, revise the parameters of that class and go to step 8. No: put the sample in hold and go to step 14.

Step 14. Has the number of samples in hold reached M? No: go to step 12. Yes: go to step 15.

Step 15. Calculate parameters for prospective class.

Step 16. With $\overline{x}_k$, s_k^2, etc. from step 15, calculate a value of A in equation (54) for each of the M samples by using the values of x_k for the particular sample in equation (54) with the minus sign. Does the largest value of A satisfy $A < A_o$? Yes: go to step 17. No: discard the first one of the M samples held for step 15 and go to step 12.

Step 17. Designate a new class with the parameter values which are extant (from step 15) and the mean $\overline{A}$ of the sample values of A.

Step 18. Empty the hold from step 14 and go to step 8.

Step 19. Subtract one from the value retained for the P parameter and retain the new value. Is the result less than one? Yes: go to step 20. No: go to step 25.

Step 20. Is the smallest A_{ij} less than A_F? Yes: go to step 21. No: go to step 22.

Step 21. Combine the classes i and j, compute the parameters (including $\overline{A}$) of the resulting class k and the parameters $A_{k\ell}$ relating it to each other class ℓ, and go to step 20.

Step 22. Prepare a print-out/read-in tape with re-start versatility.

Step 23. Print out the classification map, the class pair parameters A_{ij} and for each class the parameters m, $\bar{x}_k$, s_k, $\lambda_{k\ell}$, and $\bar{A}$ including all channels k and pairs of channels k and ℓ. Identify the print-out.

Step 24. Stop.

Step 25. Is this a re-start run? No: go to step 28. Yes: read revised control parameters W_{max}, A_F, P, and A_R and the re-start tape (of step 22) and go to step 26.

Step 26. Is W greater than W_{max}? No: go to step 28. Yes: go to step 27.

Step 27. Combine the pair of classes i and j, which correspond to the smallest A_{ij}, into a single class k, compute the class parameters and the $A_{k\ell}$ which relate it to each other class ℓ, and go to step 26.

Step 28. The extant membership of the establish classes resulting from the completed pass, the re-start after a prior classification or standardized pre-classification, give parameter values m, $\bar{x}_k$, s_k, $\mu_{\Sigma F}$, and $\sigma_{\Sigma F}$ (for equation (54)) which retain throughout a new complete pass of the data (to be revised only at the end of the data pass) for a revised classification.

Step 29. Read the upcoming sample of data in the most economical order (e.g., first, second, ...).

Step 30. Use the values of x_k for the sample and the plus sign in equation (54) to calculate a value of A for each of the W classes.

Step 31. Classify the sample by the class with the smallest A, which value remember (for step 35).

Step 32. Does the program reach the end of the sample sequence? Yes: go to step 33. No: go to step 29.

Step 33. The W established classes now have new memberships but parameter values from the previous classification. Does the smallest class i now have less than six members? Yes: go to step 34. No: go to step 35.

Step 34. The smallest class i has some smallest A_{ij} identifying its closest neighbor class j. Is A_{ij} less than A_R? Yes: Combine classes i and j and go to step 33. No: hold class i for step 35 and go to step 33.

Step 35. By the new memberships, revise the parameters for all classes with not less than six members and revise all A_{ij} for which both classes i and j are not less than six. Those classes with less than six members must retain a value of six for m for any next classification. Go to step 19.

Charles C. Dalton
Flight Data Statistics Office
Aerospace Environment Division
Aero-Astrodynamics Laboratory

APPROVAL:

William W. Vaughan, Chief
Aerospace Environment Division

REFERENCES

1. Dalton, C. C., "Notes for the Improvement of a Remote Sensing Multispectral Data Non-Supervised Classification and Mapping Technique," NASA Technical Memorandum TM X-64762, July 27, 1973.

2. Jayroe, R. R., "Unsupervised Spatial Clustering with Spectral Discrimination," NASA Technical Note D-7312, May 1973.

3. Su, M. Y., "The Composite Sequential Clustering Technique for Analysis of Multispectral Scanner Data," NASA Contractor Report CR-128999, October 1972.

4. Dalton, C. C., "Request for Program of Algorithm from NASA TM X-64762 and MSFC Memo S&E-AERO-YF-2-73 on Account No. 177-32-71 (Task Assignment J99)," MSFC Memo S&E-AERO-YF-3-73, September 19, 1973.

5. Burington, R. S., and May, D. C., Jr., "Handbook of Probability and Statistics with Tables," Handbook Publishers, Inc., Sandusky, Ohio, 1953.

APPROVAL

NOTES FOR THE IMPROVEMENT OF THE SPATIAL AND
SPECTRAL DATA CLASSIFICATION METHOD

By Charles C. Dalton

The information in this report has been reviewed for security classification. Review of any information concerning Department of Defense or Atomic Energy Commission programs has been made by the MSFC Security Classification Officer. This report, in its entirety, has been determined to be unclassified.

This document has also been reviewed and approved for technical accuracy.

W. W. VAUGHAN
Chief, Aerospace Environment Division

for L. D. RICHARD
Acting Director, Aero-Astrodynamics Laboratory